Lecture 7: Inter-Domain Routing
Credits: Based on lecture by Rob Sherwood

What

- Last time: Intra-domain routing protocols (IGP)
  - Last time
  - OSPF link state
  - RIP distance vector
- Today: Inter-domain routing protocols (EGP)
  - Border Gateway Protocol v4
  - Path vector routing protocol: list possible paths
  - No other EGP's today... why?

Why Inter vs. Intra?

- Why not just use OSPF everywhere?
  - E.g., hierarchies of OSPF areas
  - Hint: scaling is not the only limitation

- Why Study BGP?
  - Critical protocol: makes the Internet run
    - Only widely deployed EGP
  - Active area of problems!
    - Efficiency
    - Cogent vs. Level3: Internet partition
    - Pakistan accidentally took down YouTube
    - Spammers use prefix hijacking

- Outline
  - History (very briefly!)
  - Function
  - Properties
  - Policies
  - Example
  - Problems and proposed solutions
History

- Why border gateway protocol?
- Historical distinction:
  - 1989: BGPv1, “directional” routing [RFC 1105]
  - 1990: BGPv2, bunch of incompatible changes [RFC 1163]
  - 1994: BGPv4 (proposed) [RFC 1654]
  - 1995: BGPv4 (actual), w. CIDR support [RFC 1771]
  - Latest revision of BGPv4 spec [RFC 4271]
- Additional information:
  - Application of BGP in Internet [RFC 1772]
  - Experience w. BGPv4 [RFC 1773]
  - Protocol analysis [RFC 1774]

High Level

- Recall notion of Autonomous System (AS)
  - Organizations that participate in EGP
  - Assigned AS Number, originally 16 bits, now 32 [RFC 4893]
- Abstract each AS down to a single node
- Exchange prefix-reachability with all neighbors
- “I can reach prefix 171.67.0.0/14 through ASes 15444 3549 174 46749 32”
- Select a single path by routing policy
- Critical: learn many paths, propagate only one!
  - Add your ASN to advertised paths

BGP State

- BGP speaker conceptually maintains 3 sets of state
  - Adj-RIBs-In
    - Stands for “Adjacent Routing Information Base, Incoming”
    - Has unprocessed routes learned from other BGP speakers
    - Contains both reachable and unreachable routes (in case later become reachable and can be added to Loc-RIB)
  - Loc-RIB (Local RIB)
    - Contains routes from Adj-RIBs-In selected by policy
    - First hop of each route must be reachable by IGP or static route
  - Adj-RIBs-Out (Adj-RIBs, Outgoing)
    - Subset of Loc-RIB to be advertised to peer speakers

BGP Example

Only 1 Router Per AS (for now)
BGP Implications

- Explicit AS path == loop free!
  - Except under churn, IGP/EGP mismatch, etc.
- Not all ASes know all paths
- AS abstraction loss of efficiency
- Shortest AS path not guaranteed
- Scaling
  - 32K ASes
  - 300K+ prefixes

BGP protocol details

- Border routers must connect over TCP port 179
  - Bidirectionally exchange messages over long-lived connection
- Base protocol has four message types
  - OPEN – Initialize connection. Identifies BGP peers and must be first message sent in each direction
  - UPDATE – Announce routing changes (most important msg)
  - NOTIFICATION – Announce error when closing connection
  - KEEPALIVE – Make sure peer is alive
- Extensions can define more message types
  - E.g., ROUTE-REFRESH [RFC 2918]

Anatomy of an UPDATE

- Withdrawn routes: List of withdrawn IP prefixes
- Network Layer Reachability Information (NLRI)
  - List of IP prefixes to which path attributes apply
- Path attributes – various info. about NLRI
  - ORIGIN, AS_PATH, NEXT_HOP, MULTI_EXIT_DISC, LOCAL_PREF, ATOMIC_AGGREGATE, AGGREGATOR, …
  - Each attribute has 1-byte type, and 1-byte flags, plus length
  - Can introduce new types of path attribute—e.g., used AS4_PATH for 32-bit AS numbers

Transport Details

- OPEN msg negotiates capabilities [RFC 3392]
  - E.g., to advertise support for AS4_PATH
- A full information exchange after connection is expensive!
  - Keep connection open indefinitely to exchange periodic updates
- Session resets are expensive (both in CPU and to the entire network!) and should be avoided.
Advertisements

- NLRI: 171.67.0.0/14
- AS Path: ASN 15444 3549 174 46749 32
- Next Hop IP: just like in RIPv2
- Knobs for traffic engineering
  - Metric, Weight, LocalPath, MED, Communities
  - Lots of voodoo

Getting Your Hands Dirty

- RouteViews Project: http://www.routeviews.org/
  - telnet route-views.linx.routeviews.org
  - show ip bgp 171.67.0.0/14 longer-prefixes
- Note that all paths are learned internally
- Not a production device

2-minute stretch

Route Selection 1/2

- Next-Hop reachable?
- Prefer highest weight
  - Computed using some AS-specific local policy
- Prefer highest local-pref
- Prefer locally originated routes
- Prefer routes with shortest AS path length

Route Selection 2/2

- Prefer path with lowest origin type
- Prefer route with lowest MED value
  - But note can only compare MEDs from same AS
- Prefer eBGP over iBGP
- Prefer routes with lowest cost to egress point
  - Hot-potato routing
- Tie-breaking rules
  - E.g., lowest router-id, oldest route

External vs. Internal BGP
External vs. Internal BGP

Customer/Provider AS relationships
- Customers pay for connectivity
  - E.g., Stanford pays Cogent
- Customer is a stub, provider is a transit
  - Amount and cost structure can vary wildly
- Many customers are multi-homed
  - Stanford also connects to Calren/Internet2
- Typical policy: prefer routes from customers

Peer relationships
- ASes agree to exchange traffic for free
  - Penalties/renegotiate if imbalance
- Tier 1 ISPs have no default route: all peer with each other
- You are Tier \( i + 1 \) if you have a default route to a Tier \( i \)

BGP Relationship Drama
- Cogent vs. Level3
- Level3 and Cogent were peers
- In 2005, Level3 decided to start charging Cogent
- Cogent said No
- Internet partition: Cogent’s customers couldn’t get to Level3’s customers and vice versa
  - Other ISPs were affected as well
- They came to a new, undisclosed agreement 3 weeks later

BGP Security
- Anyone can source a prefix announcement
  - BGP is not very secure
- YouTube’s prefix is 208.65.152.0/22
- Pakistani government ordered YouTube blocked
  - PieNET advertised 208.65.152.0/23 and 208.65.152.128/23
  - Longest prefix match caused world-wide outage
- Spammers steal unused IP space to hide [Feamster]
  - Advertise very short prefixes—why?
- Secure BGP is currently in the works

BGP Problems and Solutions
- Security
- Convergence
- Scaling (route reflectors)
- Traffic engineering – AS preprending
- Multiple stable solutions – BGP “Wedgies”
BGP Convergence

- Given a change, how long until the network re-stabilizes?
  - …depends on the change: sometimes never.
  - Open research problem: “tweak and pray”
  - Distributed setting is challenging
- Easier: Does there exist a stable configuration?
  - Distributed: open research problem
  - Centralized: NP-Complete problem! [Griffin’99]

Scaling iBGP: Route Reflectors

iBGP Mesh == O(n^2) mess

Solution: Route Reflectors O(n^k)

Traffic Engineering

- “Route-map” programs to set weights
- Route filtering: input and output
- More specific routes: longest prefix
- AS prepending: “32 32 32 32”
- Imprecise science

BGP Wedgies [RFC 4264]

- A Common config:
  - Prefer customer routes over non-customer
  - Then prefer shortest AS path
1.2.0.0/16: AS 1

AS 3 prefers routes from AS 2